Future-Proofing with AI: Improving Decision-Making for the Future through AI Model Building and Stakeholder Deliberation

ANONYMOUS AUTHOR(S)*

Research exploring how to support or improve human decision-making through historical data has often taken the form of using machine learning to automate or assist human decision-making systems. We take an alternative approach for improving human decision-making, using machine learning to help stakeholders surface ways to "future-proof"—improve and make fairer—decision-making processes. We created *Future-Proofing with AI*, a tool that enables people to examine strengths and shortcomings of decision-making in the past and deliberate on how to improve future decisions by creating and evaluating machine learning models. We apply this tool to a context of people selection, having stakeholders—decision-makers (faculty) and decision-subjects (students)—use the tool to improve graduate school admission decisions. We report the interactions and discourse during user studies, describing how participants used the tool to discern patterns in past decision-making and engaged in deliberation to reach common understanding of admissions decision-making. Our work offers a new application for human AI collaboration and advances efforts of using machine learning to support human decision-making by illuminating: i) ways in which the tool was helpful to stakeholders as well as ideas for the future, ii) challenges and considerations to bear in mind when future-proofing decision-making with different stakeholder types, and iii) insights on expanding community-engaged participatory AI design.

CCS Concepts: • **Human-centered computing → Human computer interaction (HCI)**.

Additional Key Words and Phrases: Deliberation, Participatory algorithm design, Organizational decision-making

ACM Reference Format:

Anonymous Author(s). 2018. Future-Proofing with AI: Improving Decision-Making for the Future through AI Model Building and Stakeholder Deliberation. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 20 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Past research has shown the effectiveness and advantages of using statistics and algorithms to make decisions—compared to humans, these systematic methods can replicate outcomes on the same data and in many cases produce similar or better accuracies [2, 11, 17]. Drawing on this potential and in conjunction with advances in machine learning (ML) and artificial intelligence (AI), many systems are being developed to make more effective decisions at scale in both the public and private sectors—from predicting risk for homelessness [43] and child maltreatment [10] to managing work forces [6, 41] and allocating resources such as donations or vaccines [25, 32]. Many of these systems either automate decision-making or present humans with recommendations at the time of decision-making. Despite the promise and initial evidence, research has also shown that AI and ML models can result in biased or unfair decisions and effect harm on communities. Addressing these issues, an active area of research investigates ways to create models to be less biased, fairer, equitable, and more accountable to the community. These efforts include work around i) understanding how to design fair ML models [12, 22, 33, 44], ii) algorithmic auditing [5, 35, 45], and iii) participatory and community-engagement approaches for AI design [20, 25, 37], and iv) constructing complementary relationships between AI and humans at the time of decision-making [19, 21, 29].

We propose an alternative method aimed at improving human decision-making using machine learning. Rather than trying to improve the design or performance of algorithmic systems, we explore the use of historical data and stakeholder deliberation to future-proof—improve and make fairer—human decision-making. Analyzing historical data can help reveal patterns of successes—a premise that many ML-decision systems are based on—but also missteps and weaknesses such as human or systemic biases, non-inclusive practices and classifications, and lack of diverse representation. An ML model can externalize these patterns by learning from historical data to make predictions and displaying the results to humans. Pairing models with reflection—to help people realize behaviors to sustain or change—and deliberation—to help groups share perspectives and/or make thoughtful decisions—can surface ways to improve their decision processes, and enable people to imagine ways to leverage ML models for assisting their process based on realistic understandings of such models' benefits and risks.

As a first step toward this goal, we created a web-tool, *Future-Proofing with AI*, that helps stakeholders examine strengths and shortcomings of decision-making in the past and deliberate on how to improve future decisions by creating and evaluating ML models. We apply this tool to a context of people selection, having faculty and students use it to improve master's admissions decisions. We choose to use the master's admissions review process as predictive analytics are increasingly being used for college admissions to the concern of many due to the potential to inflict disparate harms on populations. Thus, we turn our attention to this topical domain to explore how to improve human decision-making. We found the tool was able to support participants to identify patterns from the data and model outcomes, explore their own preferences, and establish a common understanding with colleagues. Based on our study, we demonstrate three ways for how this tool can advance efforts of using ML to support human decision-making: i) allowing participants hands-on practice with model building to make concrete their abstract or unclear preferences around selections, ii) highlighting the similarities and differences in preferences and perceptions of fair decision-making between stakeholder types, and iii) offering insights into how participatory AI design may be advanced through the interactions with our tool, including the use of *Future-Proofing with AI* as a training or educational tool with community stakeholders at the initial stages of participatory AI design.

Our work makes contributions to emerging literature on human-centered use of AI and fair decision-making in the field of FAccT and human-computer interaction. We offer a new use case of AI, future-proofing with AI, for improved human decision-making, and demonstrates its potentials and limitations. We also provide insights on how to expand community-engaged participatory AI design and lessons for engaging with different stakeholder types.

2 RELATED WORK

We first review prior work on using AI and ML to automate or assist human decision-making to situate our approach. We then review work on participatory approaches to AI design, data-driven self-reflection and deliberation, particularly in the context of stakeholder-centered AI design, and describe how they inform our method.

2.1 Designing Fair ML Models and Decision-Making

2.1.1 Designing Fair ML Models and Decision-Making. Efforts around the use of ML to support human decision-making have often centered around advancing the creation of fair and equitable algorithms. Researchers have investigated how to interpret and reflect the preferences of stakeholders into systems to more accurately predict outcomes [42, 44]. Others have studied how AI and humans can work complementary with one another, such as having ML models learn when to defer to humans [21, 29] or designing AI analytic tools to augment human intuition [19]. In some instances, researchers have tested if ML models can make better predictions than humans, even finding that accounting for additional human unpredictability, models can outperform human decision-makers [22]. Additional work advancing fair algorithms has focused on defining ethical AI guidelines for ML developers. Here, researchers have focused on understanding the needs of practitioners when creating fair AI systems in order to inform techniques that can help them [18] such as checklists [28] or toolkits [3, 4].

2.1.2 Stakeholder Involvement in AI/ML design. One critique of automated decision-making systems is that impacted stakeholders are rarely consulted in the design of them. The lack of involvement can lead to not only harmful systems [1] but also lowered trust in individuals who perceive algorithmic unfairness [46]. In response, a line of research studies how to incorporate community members into the process, and whether community engagement or participatory approaches can lead to fair designs of ML. Researchers have investigated working with stakeholders to uncover contextualized notions of fairness [8], design participatory algorithms [25], and analyze algorithms in practice [37] in order to create ML systems that more accurately reflect stakeholder beliefs and preferred outcomes.

Yet, designing responsible ML decision-making tools or processes is a uniquely challenging task for stakeholders not be trained or versed in ML ("non-experts"). In order to assist stakeholders in understanding automated decisions or creating ML models, researchers have explored creating various tools and interfaces. Yang et al. [48] engaged with non-experts who use ML to surface design implications of ML, while [50] and [49] created data visualizations to convey algorithmic trade-offs in more understandable ways for designers and other users. Similarly, Shen et al. [40] tested different representations of Confusion Matrices to support non-experts in evaluating ML models.

These studies advance importance work around the construction of fair algorithmic decision-making, but center on the same end result—creating automated decision-making to support humans. We draw inspiration from these works, but we focus on improving human decision-making as our outcome instead, through the use of machine learning.

2.2 Supporting Reflection and Deliberation

Self-reflection and group deliberation are two ways that can help stakeholders surface insights from data and formulate individual and organizational goals and concerns.

2.2.1 Data-Driven Reflection. Self-reflection is a process to support individuals in making realizations about themselves and even changes to their behavior. Researchers have explored how to aid self-reflection through design of technologies and strategies that help individuals collect and probe their own data [9, 24, 27]. Lee et al. [24] observed how a reflective strategy (e.g., reflective questions) increased participants' motivation to set higher goals. Choe et al. [9] found that data visualization supports helped participants recall past behaviors and generate new questions about their behaviors to explore in the data. In our tool, we purposefully include designs to support self-reflection, such as reflective questions and data visualizations, so participants can more critically analyze their personal decision-making preferences and past data for insights.

2.2.2 Deliberation. Group deliberation can be traced back to public deliberation or deliberative democracy, where citizens gather to thoughtfully discuss policies that will impact them [15]. By bringing together different stakeholder types, deliberation can help groups to share perspectives, discuss disagreements, and make thoughtful decisions. Deliberation has been pursued in the implementation of technological tools through mediums such as chatrooms, forums, and synchronous video sessions. These works demonstrate how deliberation can help increase the accuracy of crowdworking tasks ([7, 13, 39]), improve perceptions of procedural justice [14], and support building consensus amongst participants [26, 38, 44, 47]. Data-driven and organizational deliberation have been less explored for supporting participatory algorithm design, though an exception is WeBuildAI [25] which incorporates stakeholder deliberation in ML model design. As part of our sessions with *Future-Proofing With AI*, we incorporate deliberation so that participants can share opinions and more deeply consider organizational decision-making preferences and goals.

3 FUTURE-PROOFING WITH AI

3.1 Designing for Future-Proofing

Future-Proofing refers to "the process of anticipating the future and developing methods of minimizing the negative effects while taking advantage of the positive effects of shocks and stresses due to future events" [36]. It has previously been associated with sectors including technology, infrastructure [31], energy [16], and business [30] with regards to building systems that will be resilient to as-yet unknown needs or trends. For example, future-proofing a business may refer to being more data-driven, investing in the workforce, and staying on top of the evolving digital age. We interpret "future-proofing" human decision-making as creating or improving systems to be fair, inclusive, and effective for present and future communities.

With the goal of future-proofing, we designed our web-tool to enable stakeholders to build and evaluate an ML model using data that captures historical decision patterns in their organization. This tool walks people through the primary stages of building a machine learning model: 1) Data Exploration, 2) Feature Selection, 3) Model Training, 4) Model Evaluation. (See Figure 1 for the web-tool and session flow.). While it can be used by an individual user, this tool is designed to be used by multiple people who are in a same decision-maker role, or those who are affected by the decisions, so that the resulting recommendations reflect multiple people's perspectives.

In each step, there are reflection questions and deliberation opportunities to help people i) explore past data to identify possible patterns and/or surface their own or institutional biases, ii) deliberate with colleagues to come to a common ground over what future-proofing should accomplish, and iii) generate ideas around how to future-proof and/or incorporate AI/ML-driven decisions to improve decisions. To make the tool comprehensible for users of varying algorithmic knowledge, we drew inspiration for our overall tool interface from the educational coding platform Datacamp[1] whose interface combines explanation side by side with exercises.

3.2 Tool Overview

The tool begins with an overview to explain AI and ML and introduce future-proofing. We emphasize that the goal is not to create a perfect model for automated decision-making, but to identify gaps and patterns and imagine other ways to leverage ML to improve future decisions.

3.2.1 Data Exploration. The tool first allows users to examine past data (i.e., input and decision outcomes) and shows a predictive model trained on the data (AllFeaturesModel) which shows factors that played a role in outcomes and to what degree. This design is to help them understand and reflect past decision patterns. To assist this process, the tool asks them reflective questions around admissions. To encourage users to explore and become familiar with the dataset but not be overwhelmed by data, we borrow elements from the data repository Kaggle[2] to display features in a digestible way (see Figure 1).

[1] https://www.datacamp.com/

[2] https://www.kaggle.com/

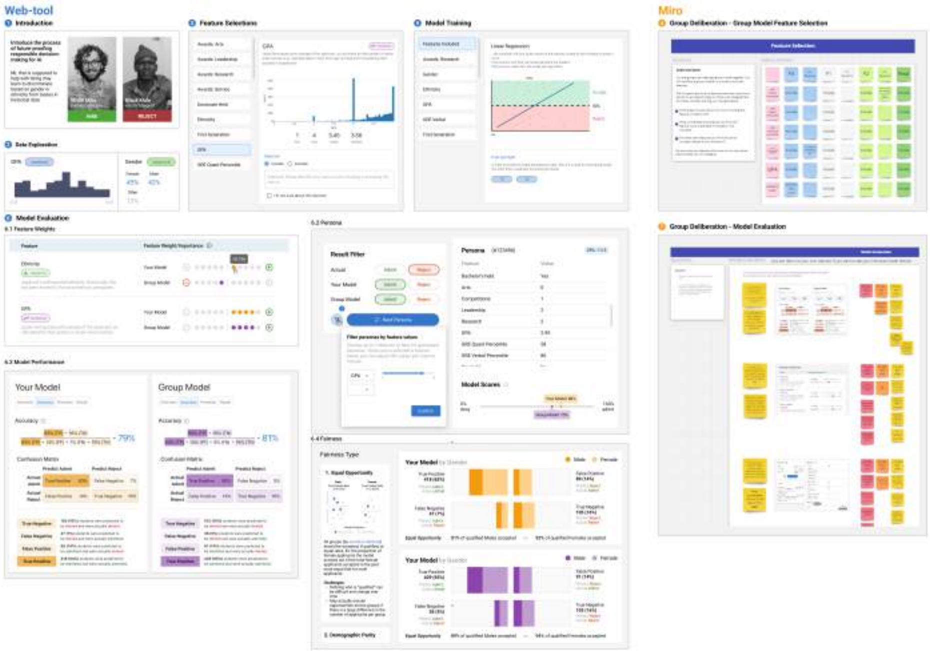


Fig. 1. *Future-Proofing with AI* and Session Flow. Content of the web-tool components are to the left, numbered in the order that the participants worked through them. Screen captures of the MIRO boards used in deliberation are to the right.

3.2.2 Feature Selection. Users can then select what factor, or features, they believe should be used, regardless of whether those features played a role in the AllFeaturesModel. This design helps users externalize their ideal decision criteria by having them choose to include or exclude each feature, share their reasoning, and/or flag if they are unsure. We ask for their reasoning and whether they are unsure on each feature to encourage self-reflection and provide material for group deliberation later. Users can also explore the dataset further by reviewing bivariate distributions between pairs of features to help generate hypotheses to explore later. For example, a user can select the feature *Ethnicity* and *GPA* to view box plots of GPA by ethnicity. In deliberation, each user's decision (include or exclude) per feature, their reasons, and whether they are unsure are displayed on a MIRO[3] board to provide a basis for deliberation; through deliberation, users finalize a set of features to include in the group model.

3.2.3 Model Training. The features that individuals select are used to train Individual Models; those the group selects through deliberation are used to train a Group Model. While models are trained, users watch a video to get a basic understanding of ML model training. We simplify the details while retaining the core steps to balance not overwhelming the user with complex concepts while providing sufficient information of how a model is trained.

3.2.4 Model Evaluation. The tool provides users with multiple ways of evaluating the models, so that they can understand what factors ML models use, kinds of errors ML models can make, and how fairness can be conceptualized in ML. This is to provide people with an understanding of ML capability, risk, and associated tradeoffs, so that they can imagine ways to leverage ML to strengthen future decision-making if they desire. On each page, a reflective question

[3] https://miro.com/

prompts users to think about the costs and benefits of human vs. automated decision-making. For example, on Personas, the question is, "If the prediction from the models and the actual admission decision differs, which do you agree with and why?" For each page, group deliberations are also held for participants to share answers to the reflective question and reactions to the model and tool activities.

Feature Weights. The tool displays the feature weights of the individual vs. group model. If a user did not select a feature that the group did or vice versa, no weight will be shown for the corresponding model.

Personas. To provide a tangible idea on who is receiving correct vs erroneous predictions, the tool allows people to retrieve personas—profiles of anonymized applicants displaying their values for all features. Users can retrieve personas that were rejected or admitted by their ML models and/or past committees; matching applicant profiles are displayed with a score and model confidence; filtering allows users to narrow down the results by features such as gender, ethnicity, GPA range, etc. This allows people to probe further to see whether certain groups of applicants are impacted by the decisions.

Model Performance. Users can see metrics (e.g., accuracy, recall) for their individual vs. group model. We include a contextualized confusion matrix similar to [40] to help users understand terms like “False Positives” in the context of admissions, augmented with textual explanations inspired by [50]. We also apply visual changes to the display so if the user selects to see how a metric is calculated, only the relevant quadrants of the confusion matrix remain on the page.

Fairness. We provide two definitions of mathematical fairness—equal opportunity and demographic parity—as an introduction for users to consider what fair or unfair model outcomes look like in ML. We include graphs based on confusion matrix visualizations of [50] to demonstrate whether disparities exist in group treatment under different fairness definitions.

3.2.5 Implementation. The web-tool is built with React and Material-UI on the frontend, and Flask and MongoDB in the backend. The plotly.js package is used for displaying graphs to users. In addition to the user functions of the main web-tool, an “Admin View” provides functions for facilitating the session (e.g., downloading a flat file of all user feature selections), abstracting the needs of facilitators from highly technical tasks such as dropping a database or creating a JSON file from a JavaScript object. For group deliberation, we use MIRO boards to take advantage of existing technologies designed for virtual collaboration.

3.3 Dataset and ML Model Training

We apply this web-tool in the context of the master’s admissions review process of a specific discipline at a public university. We submitted an IRB that allowed us access to historical admissions data, and we worked with the graduate school office to obtain anonymized applicant data with admission decisions. We were unable to obtain letters of recommendation or essays for privacy and thus could not include them. We removed incomplete or pending applications, anyone who was not a master’s applicant, and applicants with GRE scores before scoring changes in 2011. The final dataset consisted of 2207 applicants, from Fall 2013 to Fall 2019 cycles. Next, we removed irrelevant columns (e.g., timestamps) and engineered features such as using parental education to construct *First Generation*, or having researchers code features using external knowledge, such as *Tier of Undergrad Inst.* (see full details on Table 2 in the Appendix.) Although gender and ethnicity are not permissible for use in admissions decisions, we included them for discussion purposes. This resulted in the 18 features shown in Table 1.

Given the likelihood that our participants would be of varying ML knowledge levels, we used a simple machine learning model. We tested training different models (i.e., Decision Tree, Ridge Regression, Lasso Regression, and

Linear Regression) with our dataset, achieving a similar baseline accuracy of 75%-80% for all models when using all 18 features. We chose to use linear regression with a 70/30 train-test split because it is fast to train, provides us with easy-to-understand insight into the model through feature coefficients, and is comparatively easier to explain to participants than more complex models. The fast training time allowed us to create models during sessions in real time, and the feature coefficients and straightforward nature of the model made it a preferable choice to use with participants.

3.4 Method

3.4.1 Participants. In order to explore how *Future-Proofing with AI* can affect fair and responsible human decision-making, we conducted group study sessions with nine student participants (decision-subjects) and seven faculty participants (decision-makers) at a public university. Historically, master's admissions review committees have consisted solely of faculty members. Via email, we recruited those with experience on the committee and those without for varying perspectives. Although students do not currently serve on the review committee, we felt it was important to include the perspectives of impacted stakeholders. We recruited master's students by posting a sign-up with details to a general Discord channel and Pride Discord channel for current students. We held separate sessions for faculty and students so as to not exacerbate the power differentials between the groups, especially given the possibility that some students may have taken or may take a faculty participant's course in the future.

To protect the identities of the participants, we only provide aggregate statistics. Participants' ages ranged from 18-74, and 62.5% identified as female (37.5% as male). Three are of Hispanic, Latino, or Spanish origin. Ten described their ethnicity as white, three as Asian, one as American Indian or Alaska Native, one as Black or African American, and one answered "Other-Jewish Tejana, mixed race". Nine participants (56.25%) reported a Bachelor's degree as their highest degree completed at the time of the study, one (6.25%) reported a master's degree, and six (37.5%) reported a Doctorate. We also asked the participants about their current knowledge on programming and computational algorithms [23]. The average participant programming knowledge reported was 2.5 (between 2-"A little knowledge-I know basic concepts in programming" and 3-"Some knowledge-I have coded a few programs before"). The average participant computational algorithm knowledge was 2.1 (between 2-"A little knowledge-I know basic concepts in algorithms" and 3-"Some knowledge-I have used algorithms before").

3.4.2 Group Session Procedure. We conducted four virtual Zoom sessions: two with faculty members and two with students. The participants interacted with the tool to create and evaluate admissions decision-making models. Sessions lasted 2-2.5 hours each, with 3-5 participants each, and participants were compensated with $80 Amazon gift cards for their time. Group segments took place in the main room; individual segments were held via breakout rooms where each participant worked on the tool with a facilitator who prompted think-aloud or interview questions so that participants could describe experiences they may not have wished to share in a group setting.

The facilitator began with an overview of the activity while participants opened the web-tool on their computers. In the sessions[4], the facilitator displayed the feature weights of the AllFeaturesModel, a linear regression model trained with all 18 features of the dataset to help participants get familiar with working with ML models. Participants then worked through Data Exploration. Next, participants worked on Feature Selections for their individual model in breakout rooms. Everyone returned to the main room to discuss features selections for the group model. The facilitator exported the participant feature selections into a shared MIRO board (see Figure 3) and guided participants through deliberation.

[4] The first two sessions began with Data Exploration. Upon reflection from session debriefs, the facilitator modified the remaining two sessions so that displaying the AllFeaturesModel preceded this

Due to timing, the decision to include a feature was based on the majority of votes once discussion abated or if no noticeable agreement was observed from ongoing conversations. Although not ideal, the primary facilitator acted as a tiebreaker to ensure enough time for the remaining activities. Once completed, the group's feature selections were input to the tool, and all models were trained. Everyone watched the Model Training video together before moving onto Model Evaluation as a group. Then participants explored the Personas page of Model Evaluation individually with their facilitator guiding them through how to use the page so as to test the tool without distractions. Everyone returned in the main room to share the personas they explored or patterns they observed. The session concluded with wrap-up interviews in breakout rooms and an exit survey on Qualtrics.

3.4.3 Analysis. All sessions were screen-recorded using the Zoom recording feature. Recordings were transcribed in Otter.ai with errors fixed manually. Using the qualitative data analysis method [34], the primary researcher reviewed transcripts and MIRO board text, and generated initial codes based on the tool components and questions asked during sessions; and the codes were then further synthesized.

4 FINDINGS

Our participants were thoughtful in their reflections and discussions, indicating genuine interest in the subject matter and explorations of the tool. Below, we detail how participants interacted with the session components, their reactions and interactions with the activities. Faculty participants are denoted with an F-prefix, and students with a P.

4.1 Identifying Prospective Applicants

We asked participants[5] to describe what matters to them in applicants and the overall admitted class for the master's program to help ground their thinking about admissions from a holistic perspective. A common thread in nearly everyone's answer about what mattered to them in an incoming class was diversity. This was defined by some as diversity in race and ethnicity (P5, F1, F2), gender (F1, F2), age (P5, F1), socioeconomic background (F1), work background (P5, P6), and general background and experiences (P1, P2, P3, P4, P7, P8, F3). P8 emphasized the importance of a class that mirrored the real world: "especially since we're going to be working a lot in groups, it would be nice to have like a pretty good representative of, I guess, just the world around us." Student participants also shared a preference for incoming students to have a collaborative nature. As P1 described, he hoped to see "students from different backgrounds who are able to work with people from different backgrounds themselves", perhaps reflective of coursework or work experiences requiring group work.

Participants mentioned a few quantifiable, academic-related qualities ("education background" -F5, "high grades and GPAs" -F1, "rating of college from which app[licant] has a degree" -F2), the vast majority of applicant attributes were abstract and harder to define or measure. Participants highlighted that applicants needed to exhibit a clear purpose for grad school (P7, F3, F1, P1, P9), be passionate or excited to learn (P2, P3, F3), and be "other-oriented" or service-driven (F2). This overwhelming emphasis on conceptual qualities indicates the nuances in human decision-making and potential challenges a tool may have in assisting or improving human decision-making.

4.2 Data Exploration and the "AllFeaturesModel"

The AllFeaturesModel assisted participants in identifying patterns or trends from past decision-making. Students were surprised at some feature weights, such as a small but negative weight for *Awards: Research* and a comparatively strong

[5] 1 of the 4 groups did not discuss these questions due to time constraints.

positive weight for *Work Experience*. P7 observed the weights of the 3 GRE components were all positive and higher than the weight of *Tier of Undergrad Inst.* and shared her surprise and conjectures for how features may be correlated: "I'm actually really surprised at how much the GRE is taken into account...so I know that the Tier of Undergraduate Institution is kind of largely determined by your, your class background...so I'm like, I'm wondering if like the GRE weight is meant to like, balance that out." This point was revisited later in the session as another participant shared how they perceived GREs balancing out a low GPA or a low tiered school.

Faculty participants displayed some but less surprise over the weights, perhaps because the majority had previously served on an admissions committee and felt the model was reflective of the patterns they'd observed. But similar to P7, F3 observed the positive weights on the GRE features. She suggested an interpretation about what past data may indicate about decision-makers: "We're not using GRE anymore, but it points out the reliance on the GRE in past decisions." She also observed how these patterns could inform how to advise prospective applicants after seeing the comparatively high weight the model placed on work experience: "It's almost as if you were to recommend to a student based again, on past [admission's] experience, how best to prepare for a master's degree, it would be to get some work experience."

Participants generated new ideas for analyzing applicant data as they assessed the AllFeaturesModel. P6 and P9 wondered whether feature weights would vary if broken out by additional criteria such as degree concentration or work experience field. F7 asked whether separate models should be considered for domestic versus international students given the difference in admissions prerequisites for the two (e.g., requirement of TOEFL scores for international students). F5 wanted to see information about an applicant's past institutions in a more comprehensive manner than *Tier of Undergrad Inst.* explaining, "I'd really like to know public schools versus private schools, and size of institution," expounding later that she valued the experiences of students who come from regional or community colleges.

4.3 Feature Selection

Participants sometimes provided personal experience and beliefs behind including or excluding features, highlighting their diverse backgrounds and experiences. P2 explained that since she did not have to submit GRE scores, she did not think it was necessary especially when through essays, reviewers could "see their story rather than just a number". Conversely, P5 included GRE features because as a first generation student, she worked multiple jobs to support herself during school, affecting her GPA, thus "for me, it was like important to submit my [GRE] score so that the admissions committee can see, hey, this person has a 2.8 GPA, but 8 years later...they've taken this test and like are a competent reader and math person and writer." Similarly, F7 shared that growing up in a country that strongly stressed testing had conditioned him to feel the same around including GRE scores.

A handful of participants excluded features based on the limitations they believed an ML model would have when interpreting it. F6 explained that he would personally use descriptive information about awards for making decisions but that the quantified features were insufficient for a model. P7 omitted *Awards: Arts* after seeing the data and feeling it was inadequate for a model. P3 chose not to include *Gender* based on concerns that the model would leave students out: "The only options there were, were male and female, which leaves out a bunch of people...I don't know what a model would do if someone said like they were non-binary, like, would it just automatically exclude them because it wasn't an option in the past data?"

Some participants wished to update models with custom feature weights to account for different scenarios. F3 suggested that a model handle if-then scenarios when evaluating applicants, where if the value for a feature fell in one range, the model would re-weight another feature in order to balance the first. Some student participants also expressed a desire to customize feature weights. P1 explained wanting for features "a sliding scale of like how weighted or important it is." P2 agreed, thinking how applicants with lower GPAs but harder coursework should not be penalized.

Participants sometimes included features for the nuanced context they felt the features would provide the model. This often arose when discussing gender, ethnicity, and class-related features. The feature *First Generation* and the topic of socioeconomic class were openly discussed by faculty and students for inclusion in decision-making. In fact, all but one participant chose to include it in the model. A few student participants explained they viewed *First Generation* as the closest (but "not imperfect") proxy to class, sharing concerns of how these applicants face disadvantages around affording school or navigating the application process. *Gender* and *Ethnicity* were approached differently by students and faculty. While both said decisions should not be made solely based on gender or ethnicity, students were more open to discussing the use of them in order to ensure fairness and diversity and contextualize students' experiences, whereas faculty members tended to avoid discussing the features in detail. F5 had difficulty articulating why including ethnicity was important for her. A few like F3 wanted to know the features for keeping a check of the diversity of the applicants, but that it otherwise did not factor into her decisions. F6 began sharing that gender and ethnicity should only be evaluated holistically as part of an applicant's positionality, but discussion over the two features stalled after. We do not suggest that this reflects how faculty members feel about how gender and ethnicity factor into admissions, but instead is more reflective of the rules or expectations that come with such bureaucratic tasks. A few expanded on their personal feelings during interviews, such as F2 sharing he felt the school had a ways to go in improving ethnic diversity and F4 commenting that historically ethnicity and diversity have been topics of avoidance for admissions.

4.4 Model Outcome Evaluation

4.4.1 Model Performance. Participants shared their impressions of their model performances and were asked for their thoughts on potential harms of models in terms of false positives (i.e., students admitted by the model but rejected by past committees) and false negatives (i.e., students rejected by the model but accepted by past committees).

In contrast to faculty participants who preferred low false positives, nearly all students shared a preference for models that had high false positives. We observed students displayed thoughtful inclusion-oriented reasoning, such as P1, P3, and P4 preferring a model that made more false positives, thus "erring on the side of giving people a chance" (P1). P3 felt admitting an "unqualified" candidate would not harm others but could benefit the applicant. Faculty members shared differing opinions. F3 worried about mistakenly admitting a student: "It's more detrimental to bring a student who cannot succeed. And I think that is, you know, kind of falls on the university." F5 felt that the pressing potential harm of admissions lay in lack of time or human resources for review, encouraging a pivot of thinking of "how do we help committees in the future, as opposed to reinforcing the committees of the past?"

A few participants felt that non-perfect accuracy scores could be useful signals to identify areas of decision-making for further investigation. P6 thought a non-perfect accuracy could be used identify whether specific groups of applicants are being unfairly denied: "to find out if there are...certain features of students that fall into that group that humans have denied...that the model accepts or that the humans accept that the model denied." P9 added that non-perfect accuracy scores could be used to improve the model by identifying additional features to include: "if you allow for a gap, we do allow for more of that less specific information. So you can see what was relevant that you need to add into your model in the future." F3 actually wanted to see a high accuracy on her model, however she shared similar feelings that a gap in accuracy represented an opportunity to improve admissions and the model creation process.

4.4.2 Personas. On the Personas page, participants browsed student profiles and generated hypotheses for the decision-making patterns they were observing. P5 was interested in whether past committees did in fact balance features such as GPAs and GREs. She observed that many applicants with high *GRE Quant* scores and low *GPAs* were accepted in reality. However, after she found a persona with a high GPA/low GRE who was rejected in reality but admitted by the models, she wondered whether reviewers did not actually balance low GREs with high GPAs and penalized the low score. F2 also began noticing a pattern where students accepted in reality but rejected by models had high *GRE Quant* but lower scores elsewhere, saying "I wonder why we accepted them? Looks like we overvalued the GRE quant".

Some used persona exploration to determine what context to seek out on the rest of an applicant's package. F1 noticed that the scores of one persona (high GRE, low GPA) resembled someone returning to school from working, and wanted to know more about their work experience background: "Have they been working? Maybe they've been working in a library for the last couple of years." F6 walked through a persona, suggesting possible context for the details he was viewing ("Tier 1 undergraduate institution, so probably a regional schooler"), explaining that his next step would be going into their personal essay for the full story. P6 was confused by two personas that the models rejected but the actual decisions were acceptances. Because of the lack of evidence from the features he could see on the screen for accepting them, he was left to conclude that data such as essays made a difference. He felt models could be improved by having access to the reasons behind these candidates' acceptances: "That would help to understand and improve the model better."

4.4.3 Fairness. We observed that the fairness definitions page of the tool was challenging for participants to grasp, as they struggled during think-alouds and group discussions to share their thoughts. F3 said, "I'll admit, I didn't really, yeah this one was a bit above me." This may have been because we provided two mathematical definitions of fairness, equal opportunity and demographic parity, which may not be the most intuitive way of thinking about an abstract concept as demonstrated by [42]. We observed participants found it easier to talk about attributes that contribute to fair decisions rather than discussing quantitative measures of fairness. So we shifted our objective to using Model Evaluation pages as discussion points for surfacing participant concerns and thoughts on fairness considerations for admissions decision-making.

Participants, particularly students, were outspoken about the ways they wanted to see fairness incorporated into decision-making. Students wanted to ensure applicants from traditionally marginalized backgrounds were not disadvantaged. They suggested additional features such as disabilities and veteran status (P2), a more reliable indicator for class other than *First Generation* (P3, P5, P7, F3), and improving features such as making gender more inclusive (P1, P3). P1 wanted gender to include more options than "male" and "female" in order to acknowledge the exclusion that non-binary, gender non-conforming, or trans people face in opportunities: "It has the potential to really sort of like disrupt somebody...(their) ability to participate fully in school...and all the extracurricular type stuff." P3 explained first generation students are disadvantaged compared to others when applying for higher education, as related to parental help, and should be a consideration for fair admissions practices. P7 insisted that class was an essential consideration in fair admissions, could not be substituted by *First Generation*, and needed to be included as a separate feature in the dataset.

5 DISCUSSION

We expand on our findings to discuss the implications of our results from using *Future-Proofing with AI*, namely: i) ways the tool was helpful to stakeholders as well as ideas for the future, ii) challenges and considerations to bear in mind when future-proofing decision-making with different stakeholder types, and iii) insights on expanding community-engaged participatory AI design.

5.1 Future-Proofing With AI

We discuss the ways the tool supported stakeholders and the ideas participants came up with to help humans make fairer and more effective decisions.

5.1.1 Putting Abstract into Practice. We intended *Future-Proofing with AI* as a tool to assist participants in uncovering gaps in past decisions so they could identify ways to improve human decisions. We found the tool also helped participants gain a better sense of their own decision-making patterns, understand the perspectives of others, and begin to comprehend how ML models work.

Activities in the model evaluation stage helped participants better understand their own decision-making patterns and gain insight into how features might be used by a model. For P7, the hands-on practice of reviewing personas helped her recognize differences in which features she cared about when thinking about them abstractly (in Feature Selection) compared to when she saw them as attributes of an applicant. For P1 and P3, comparing their individual and group models' feature weights help them see how a model interpreted features they were unsure about earlier. P3 decided she would have included *First Generation* in her individual model after seeing in the group model that including it did not penalize those students from being admitted.

Group deliberation was central to helping participants think critically about their own preferences, and often surface the limitations of historical data for a model as well. On a couple occasions, group deliberation led to several participants experiencing a change of heart for including or excluding a feature. Student participants in the first session were split on whether to include *Gender* and *Ethnicity* in a model though all of their justifications were based on the same goal: ensuring future classes were diverse and equal and that decision-making did not exhibit biases. After P3 pointed out concerns that the data was not representative enough of ethnicity or inclusive of non-binary gender identities, all the student participants chose to exclude these for the group model with P4 referencing the discussion later as what convinced him. Even though deliberation did not result in unanimous agreement for every feature, participants all found it helpful in identifying gaps in their perspectives and establishing an understanding about one another. Faculty members even felt this was something that review committees could benefit from.

5.1.2 Ideas from Future-Proofing. We asked participants how they thought AI tools and non-technological tools could and should assist them in decision-making. Student participants suggested support centered on assisting reviewers in identifying their potential biases. P1 suggested models from *Future-Proofing with AI* could be used to act as a "bias check", P3 expanding that reviewers could use the tool after making decisions to check their personal biases and then provide decision rationale. Faculty though, focused on implementations to help decision-makers review more effectively and efficiently, such as ML summarizing and then describing important characteristics about an applicant to assist reviewers, identifying which applicants were wildcards and the "Larry Birds" of the bunch, and nudging reviewers about pacing themselves or reviewing certain components.

Students and faculty also had different approaches when discussing non-technological implementations to improve decision-making. While students focused on improving diversity outside of the review process, most faculty members were concerned with training or review support tools. Students insisted on improving recruitment and outreach programs in order to expose a wider audience to the program earlier. P7 emphasized that "if you want more students like me, if you want more Hispanic students", outreach, mentorship, or pathway programs were necessary. P3 pointed out the importance of downstream efforts as well, once a student is accepted, and the need to expand funding so that enrollment is not limited to “just those who can afford to pay the tuition”. Their ideas were all striking reminders about how efforts of improving decision-making are not siloed to the one area of selections, and that other parts impact how outcomes ultimately play out.

Most faculty participants focused on tech-related aids, likely because they face the immediate task of reviewing hundreds of applications holistically. However, F4 shared three reflections about how non-tech based methods could improve the review process. First, she explained that dialogues around school diversity are not often held amongst faculty, but open conversations about AI, its biases, and the importance of diversity are necessary to start raising people's awareness. Second, she suggested a variation of futuring—designing futures and imagining what citizens of the future need—could be used by committees for envisioning future classes. Finally, she shared recruitment efforts must be more intentional than a checklist of schools to visit and described her own experiences of recruiting students of color by building meaningful relationships with prospective students early.

5.2 Working with Different Stakeholder Types

5.2.1 Decision-Makers vs. Decision-Subjects. Faculty (decision-makers) and students (decision-subjects) exhibited similarities and differences to their views of certain features and how decision-making should be conducted. Early on, both faculty and students expressed a shared desire for decision outcomes to reflect diversity, and this often emerged during group discussions and the ideas they came up with around future-proofing. Both faculty and students also used similar approaches for their reasoning around including features, drawing on personal experiences or choosing to include features for the context they hoped it would provide a model.

However, faculty and students differed in a few ways such as how they viewed the risks of applying automated decision-making. Students overwhelmingly believed accepting an unqualified candidate (i.e., someone who was rejected by the past committees) was less harmful while faculty members felt the opposite. Students viewed the former as giving “people a chance”, with P1 adding that he felt false positives could be a way of “correcting for the bias in the data” given potential historical biases. However, faculty believed accepting students who had previously been rejected was more harmful, F3 explaining that if the student was not successful, the fault would lie with the institution for not providing the support. For other faculty, the urgency was simply around how to support future committees given the influx of applications in recent cycles. Another factor for why faculty and students differ may be attributed to how faculty currently lack measurements to assess whether students they accepted were successful. In that sense, it may be harder for faculty to conceptualize potential harms of admission "misses", whereas students have recently been applicants and can imagine being accepted or rejected.

5.2.2 Accountability of Decision Practices by Decision Subject. These differences in how decision-makers and decision-subjects view potential harms and where they believe efforts are needed to improve decision-making present unique challenges for decision-making practices. First, the difference in how faculty and students viewed harms of admitting someone who is "unqualified" and rejecting someone who is "qualified" signifies a consequential difference in decision-making practices. Faculty took on the perspective of institutions of how "mistakes" may impact the school while students took the perspective of applicants and the effects on them. This difference in perspective around harm translates to perceptions of fairness and trustworthiness of a decision process. It raises questions around how these differences should be accounted for in the creation of models or decision practices to be considered trustworthy by decision-subjects. It also raises questions around if/how decision-makers and decision-subjects should work together to inform human decision-making.

Second, faculty and students presented different ideas for how technological or non-technological tools can help, which may indicate a potential tension over where resources should be used as they relate to admissions. For example, as related to non-technological support: is it more important to expend money and effort to ensure future classes come from as diverse of communities as possible? Or is it more important to devote resources to the immediate issue of hiring and training faculty reviewers. How should the input of both decision-makers and decision-subjects be combined so both feel supported in their goals, especially as decision-subjects may take on a different role after graduating?

5.3 Community-Engaged Participatory AI Design

While the goal of *Future-Proofing with AI* was not to help people build an algorithm for the purpose of automating decision-making, our studies offer insights into developing community-engaged, participatory AI design.

5.3.1 A Precursor to Designing AI/ML. Participatory AI design has thus far explored how to create AI models with individuals or groups. But often stakeholders involved are non-ML experts who may struggle with understanding how their decisions, such as feature selections, actually impact the outcomes of the model. To support non-ML experts, researchers have explored ways such as improving visualizations or other tools to make ML more understandable.

Based on our observations, we propose that *Future-Proofing with AI* or a similar tool can fill the role in participatory AI design of an educational precursor to the actual design of models with stakeholders for the real world. We designed the tool to follow the traditional ML model building pipeline because we wanted participants to get the full scope, albeit simplified, of how a model is created. We observed this framing and hands-on practice with models helped participants uncover ideas for how ML can assist decision-making responsibly or ways it fell short. These are important realizations for participants to make as a preparatory step to ensure thoughtful design once they engage in AI design.

We also highlight that using this tool as a precursor can help advance community-engaged participatory AI design by transforming the experience for stakeholders into a low-pressure situation. Because the goal of *Future-Proofing with AI* is to uncover patterns or nuances in preferences, the stakes of creating an infallible model are removed and participants can feel more at ease with sharing explanations (or lack there-of) about feature selections, or critiquing the models they create. As P7 commented, while hesitating over a feature selection and then deciding to include it simply out of curiosity, "this is just like a first draft!" We agree: the model does not have to be perfect to allow participants to engage with it and surface ways to improve decision processes. Integrating a component of education and practice as a part of participatory AI design can go a long way in revealing early patterns of historical data and ideas around the use of AI/ML by stakeholders or community members.

5.3.2 Reflections: Care in Design. Although all of our participants were able to complete the session and create an ML model, we stress the importance around use of this tool with care. While the methods of data exploration and deliberation that we used were able to support our participants in making sense of past data, data analysis can still be a daunting task for many, intimidating to attempt alone and in front of others.

Some of our participants shared insecurities about exploring data due to their background or participant nerves. In these instances, facilitators worked to ameliorate apprehensions by breaking down concepts or focusing on qualitative questions. P5 shared, “I’m really not familiar with working with datasets at all, so I don’t know if I’ll make any intelligent decisions.” She began to enjoy using the tool though, particularly the Personas, exclaiming as breakout rooms closed, “I want to keep working!” F2 indicated frustration at the start of Model Evaluation, explaining, "I’m not 100% sure I’m going to be able to figure it out." With facilitator guidance, he began to gain confidence as he explored the various pages. However, facilitator assistance was not always enough: F1 declined to explore parts of the tool, saying, "that makes me feel stupid", and sharing during the interview that, "this could have been like an all day thing, and I still probably wouldn't have had enough time to figure it out."

Additionally, the topic of admissions may be unsettling for participants who feel uncomfortable sharing their personal views with groups or strangers or have had negative past experiences. We attempted to reduce negative associations for participants such as interjecting when needed, redirecting questions that stirred anxiety, and reassuring them that there was no right or wrong answer. It may not be obvious though, during group discussions, if there is a discomfort. F4 shared in her debrief that when hearing a colleague explain his inclusion of GREs, "I did feel very, actually had to look away from the screen...that was very hurtful for me to hear that because it tells me that we should require...a Blackfeet person living in a rural area in Montana to do what we did." She also commented about the power dynamics that existed in the room of participants amongst faculty members given varying experience in admissions, as a faculty member, and tenure at the school. This within-group power dynamic surprised us as we had previously only been cognizant of ensuring students and faculty did not overlap given that asymmetry of power, but is another consideration for how to responsibly engage with communities and different stakeholder types in the design of participatory algorithms.

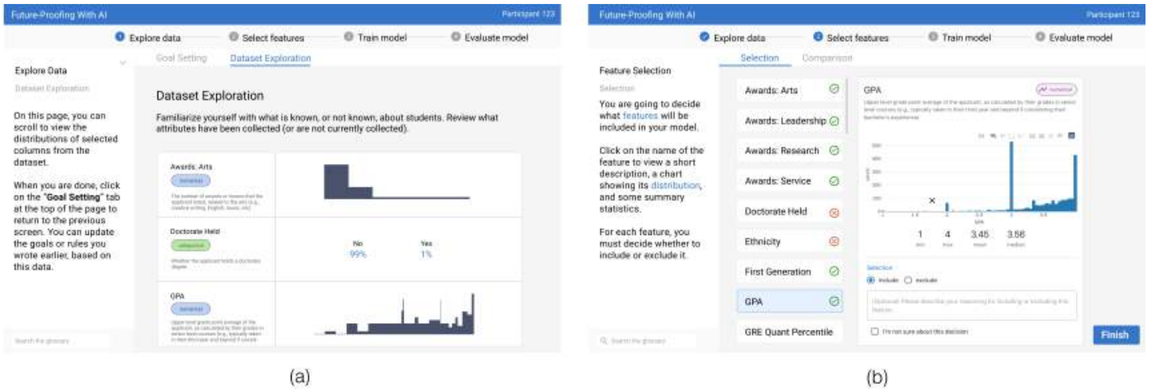
6 LIMITATIONS

Our study has limitations. Our tool used only quantitative data due to the privacy issues associated with qualitative data. We set sessions to 2-2.5 hours due to the challenges of scheduling and recognizing that longer sessions would lead to more fatigued participants. This limited the time for participants to digest and fully explore the information on the tool; splitting the session into a series can be one way to allow participants to gradually build knowledge and comfort using models in future studies. Due to the unknown variance in baseline ML knowledge of our participants and the explainability of linear regression algorithms, we use linear regression to minimize cognitive load on participants; catering our explanations to those with little ML knowledge may have made the tool and model overly simplistic to participants with greater ML experience and limited what they were able to contribute. Due to the limited time, we had to omit more sophisticated designs such as allowing a non-linear or iterative use of the tool. Future studies should explore iterative use of the tool—people changing features, retraining, and examining the models. Finally, we were only able to recruit from students who were accepted and enrolled; future studies may be done with those who have not yet but plan to apply, applied and were rejected, and were accepted but chose not to come. We intentionally held sessions for only faculty members or only students due to the inherent power dynamics between the two. Future work can explore ways to account for power dynamics in deliberation as mixing stakeholder types may affect different deliberation outcomes and group models.

7 CONCLUSION

Many important lines of research are working towards ensuring algorithmic decision-making systems are created equitably and fair. While these approaches focus on how to improve or create a machine learning or automated system, we propose using machine learning to improve human decision-making. We created a web-tool, *Future-Proofing With AI*, to explore how to support stakeholders in identifying insights about past decisions to inform the practices of human decision-makers; we demonstrate how faculty and students used the tool to explore historical data about master's applicants, engage in reflection and deliberation to create ML models, and evaluate models to discern past human decision-making patterns.

8 APPENDICES



Explore Data

Dataset Exploration

On this page, you can scroll to view the distributions of selected columns from the dataset.

When you are done, click on the **"Goal Setting"** tab at the top of the page to return to the previous screen. You can update the goals or rules you wrote earlier, based on this data.

Dataset Exploration

Familiarize yourself with what is known, or not known, about students. Review what attributes have been collected (or are not currently collected).

(a)

Feature Selection

Selection

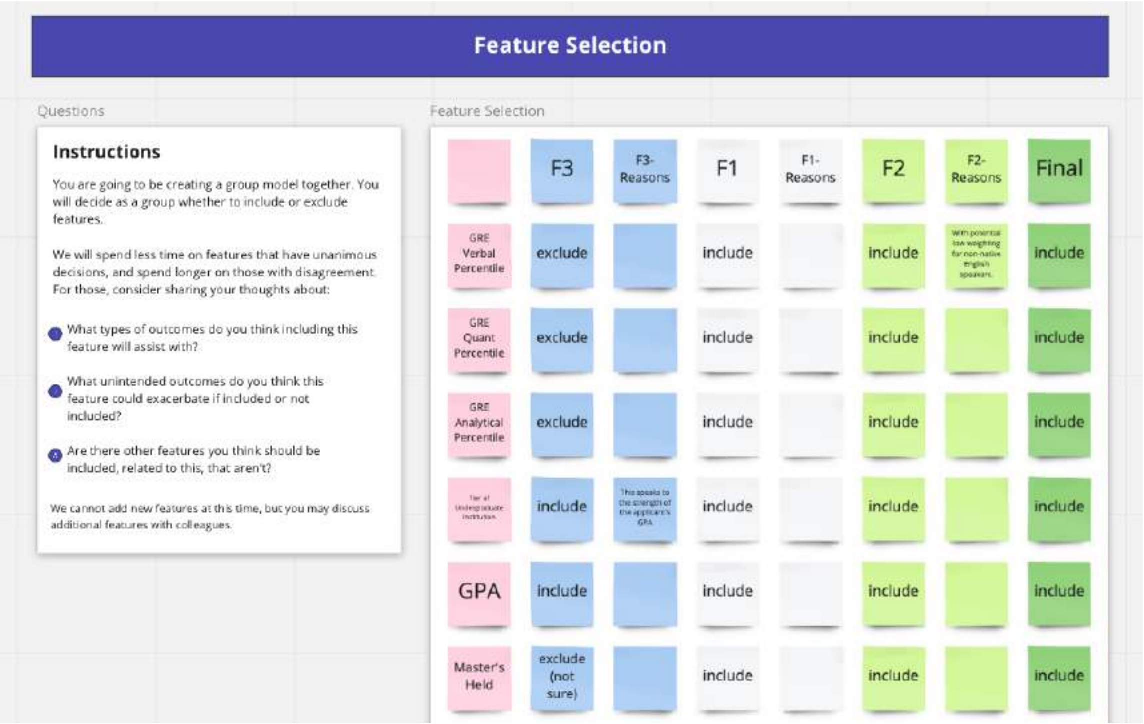
You are going to decide what features will be included in your model.

Click on the name of the feature to view a short description, a chart showing its distribution, and some summary statistics.

For each feature, you must decide whether to include or exclude it.

(b)

Fig. 2. Screen captures of the full tool interface, displaying here the (a) Data Exploration and (b) Feature Selection pages.



Instructions

You are going to be creating a group model together. You will decide as a group whether to include or exclude features.

We will spend less time on features that have unanimous decisions, and spend longer on those with disagreement. For those, consider sharing your thoughts about:

- What types of outcomes do you think including this feature will assist with?
- What unintended outcomes do you think this feature could exacerbate if included or not included?
- Are there other features you think should be included, related to this, that aren't?

We cannot add new features at this time, but you may discuss additional features with colleagues.

	F3	F3-Reasons	F1	F1-Reasons	F2	F2-Reasons	Final
GRE Verbal Percentile	exclude		include		include	With potential low weighting for non-native English speakers.	include
GRE Quant Percentile	exclude		include		include		include
GRE Analytical Percentile	exclude		include		include		include
Tier of Undergraduate Institution	include	This speaks to the strength of the applicant's GPA	include		include		include
GPA	include		include		include		include
Master's Held	exclude (not sure)		include		include		include

Fig. 3. Enlargened screen capture of a MIRO deliberation session.

Feature Name	**Description**	**Incl. by Students**	**Incl. by Faculty**
GRE Verbal %	Percentile of the applicant's GRE Verbal score.	78%	57%
GRE Quant %	Percentile of the applicant's GRE Quantitative score.	67%	57%
GRE Analytical %	Percentile of the applicant's GRE Analytic (Writing) score.	67%	57%
Tier of Undergrad Inst.	Tier 4 being top institutions and 1 being bottom. Primarily determined by aggregating several US News rankings.	44%	86%
GPA	Upper level grade point average of the applicant, as calculated by their grades in senior level courses (e.g., typically taken in their third year and beyond if considering their bachelor's experience)	89%	86%
Master's Held	Whether the applicant holds a master's degree	11%	86%
Doctorate Held	Whether the applicant holds a doctorate degree	11%	43%
Special Degree Held	Whether the applicant holds a special degree	11%	43%
Awards: Arts	The number of awards or honors that the applicant listed, related to the arts (e.g., creative writing, English, music, etc.)	78%	43%
Awards: Scholastic	The number of awards or honors that the applicant listed, related to receiving scholarships or being holding top student rankings such as valedictorian.	100%	86%
Awards: Research	The number of awards or honors that the applicant listed, related to research experience such as independent study, research grants, or writing a thesis.	89%	86%
Awards: Service	The number of awards or honors that the applicant listed, related to service or volunteering.	67%	71%
Awards: Leadership	The number of awards or honors that the applicant listed, related to holding leadership positions.	78%	57%
Awards: Competition	The number of awards or honors that the applicant listed, related to competitions or contests relevant to academia (e.g., creative writing, English, music, etc.)	78%	71%
Gender	Applicant's self-reported gender. Historically, this has been limited to two choices, male or female.	11%	57%
Ethnicity	Applicant's self-reported ethnicity. Historically, this has been limited to five race/ethnicity categories.	44%	71%
First Generation	Whether the applicant is the first in their family to receive a bachelor's degree. Inferred by the education level of an applicant's parent(s).	89%	100%
Work Experience	How many years an applicant has worked, defined by the time between the applicant's earliest work start date and most recent work end date.	100%	100%

Table 1. Set of 18 features displayed in the tool during Feature Selection as well as in the AllFeaturesModel.

Feature Names	Description
GRE Verbal %, GRE Quant %, GRE Analytical %	Directly mapped from GRE percentile columns in the dataset.
Tier of Undergrad Inst.	Mapped by taking the undergraduate institution the applicant most recently matriculated from to a tier. 4 is the highest tier, while 1 is the lowest. Tiers were formed by aggregating National, Regional, Country, and International school ranking lists from US News.
GPA	Mapped from the GRE column in the dataset that the graduate school calculated from an applicant's upper level classes.
Master's Held, Doctorate Held, Special Degree Held	Mapped as a binary from whether the applicant reported obtaining one of these degrees in their education history.
Awards: Arts, Competition, Leadership, Research, Scholastic, Service	Formed by hand-coding the 3 free-form text fields each applicant could use to report honors/awards into categories and summing these categories for each applicant.
Gender	Mapped directly from the gender column of the dataset, historically limited to only male or female.
Ethnicity	Mapped from the primary ethnicity column of the dataset.
First Generation	Formed by the education history columns of the applicant's parents/guardians. If all reported history is below a Bachelor's Degree, this value is Yes.
Work Experience	Calculated from the applicant's reported work history dates, by subtracting their earliest work start date from their most recent end date.

Table 2. Logic for how the 18 features were formed from the original dataset.

REFERENCES

[1] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias: There's software used across the country to predict future criminals. *And it's biased against blacks. ProPublica* 23 (2016), 77–91.

[2] Marc Aubreville, Christof A Bertram, Christian Marzahl, Corinne Gurtner, Martina Dettwiler, Anja Schmidt, Florian Bartenschlager, Sophie Merz, Marco Fragoso, Olivia Kershaw, et al. 2020. Deep learning algorithms out-perform veterinary pathologists in detecting the mitotically most active tumor region. *Scientific Reports* 10, 1 (2020), 1–11.

[3] R K E Bellamy, A Mojsilovic, S Nagar, K Natesan Ramamurthy, J Richards, D Saha, P Sattigeri, M Singh, K R Varshney, Y Zhang, K Dey, M Hind, S C Hoffman, S Houde, K Kannan, P Lohia, J Martino, and S Mehta. 2019. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM J. Res. Dev.* 63, 4/5 (July 2019), 4:1–4:15.

[4] S Bird, M Dudík, R Edgar, B Horn, R Lutz, V Milan, and Walker. 2020. *Fairlearn: A toolkit for assessing and improving fairness in AI. Microsoft.*

[5] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*. PMLR, 77–91.

[6] Dennis Carey and Matt Smith. 2016. How companies are using simulations, competitions, and analytics to hire. https://hbr.org/2016/04/how-companies-are-using-simulations-competitions-and-analytics-to-hire

[7] Quanze Chen, Jonathan Bragg, Lydia B Chilton, and Dan S Weld. 2019. Cicero: Multi-turn, contextual argumentation for accurate crowdsourcing. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–14.

[8] Hao-Fei Cheng, Logan Stapleton, Ruiqi Wang, Paige Bullock, Alexandra Chouldechova, Zhiwei Steven Steven Wu, and Haiyi Zhu. 2021. Soliciting Stakeholders' Fairness Notions in Child Maltreatment Predictive Systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–17.

[9] Eun Kyoung Choe, Bongshin Lee, Haining Zhu, Nathalie Henry Riche, and Dominikus Baur. 2017. Understanding self-reflection: how people reflect on personal data through visual data exploration. In *Proceedings of the 11th EAI International Conference on Pervasive Computing Technologies for Healthcare*. 173–182.

[10] Stephanie Cuccaro-Alamin, Regan Foust, Rhema Vaithianathan, and Emily Putnam-Hornstein. 2017. Risk assessment and decision making in child protective services: Predictive risk modeling in context. *Children and Youth Services Review* 79 (2017), 291–298.

[11] Robyn M Dawes, David Faust, and Paul E Meehl. 1989. Clinical versus actuarial judgment. *Science* 243, 4899 (1989), 1668–1674.

[12] Maria De-Arteaga, Riccardo Fogliato, and Alexandra Chouldechova. 2020. A case for humans-in-the-loop: Decisions in the presence of erroneous algorithmic scores. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.

[13] Ryan Drapeau, Lydia Chilton, Jonathan Bragg, and Daniel Weld. 2016. Microtalk: Using argumentation to improve crowdsourcing accuracy. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 4.

[14] Jenny Fan and Amy X Zhang. 2020. Digital juries: A civics-oriented approach to platform governance. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–14.

[15] James S Fishkin. 2002. Deliberative democracy. *The Blackwell guide to social and political philosophy* (2002), 221–238.

[16] Maria Christina Georgiadou, Theophilus Hacking, and Peter Guthrie. 2012. A conceptual framework for future-proofing the energy performance of buildings. *Energy Policy* 47 (2012), 145–155.

[17] William M Grove, David H Zald, Boyd S Lebow, Beth E Snitz, and Chad Nelson. 2000. Clinical versus mechanical prediction: a meta-analysis. *Psychological assessment* 12, 1 (2000), 19.

[18] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé, III, Miro Dudík, and Hanna Wallach. 2018. Improving fairness in machine learning systems: What do industry practitioners need? (Dec. 2018). arXiv:1812.05239 [cs.HC]

[19] Mohammad Hossein Jarrahi. 2018. Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making. *Bus. Horiz.* 61, 4 (July 2018), 577–586.

[20] Michael Katell, Meg Young, Dharma Dailey, Bernease Herman, Vivian Guetler, Aaron Tam, Corinne Bintz, Daniella Raz, and PM Krafft. 2020. Toward situated interventions for algorithmic equity: lessons from the field. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 45–55.

[21] Vijay Keswani, Matthew Lease, and Krishnaram Kenthapadi. 2021. Towards unbiased and accurate deferral to multiple experts. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (Virtual Event USA). ACM, New York, NY, USA.

[22] Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. Human decisions and machine predictions. *The quarterly journal of economics* 133, 1 (2018), 237–293.

[23] Min Kyung Lee and Su Baykal. 2017. Algorithmic mediation in group decisions: Fairness perceptions of algorithmically mediated vs. discussion-based social division. In *Proceedings of the 2017 acm conference on computer supported cooperative work and social computing*. 1035–1048.

[24] Min Kyung Lee, Junsung Kim, Jodi Forlizzi, and Sara Kiesler. 2015. Personalization revisited: a reflective approach helps people better personalize health services and motivates them to increase physical activity. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 743–754.

[25] Min Kyung Lee, Daniel Kusbit, Anson Kahng, Ji Tae Kim, Xinran Yuan, Allissa Chan, Daniel See, Ritesh Noothigattu, Siheon Lee, Alexandros Psomas, et al. 2019. WeBuildAI: Participatory framework for algorithmic governance. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–35.

[26] Sung-Chul Lee, Jaeyoon Song, Eun-Young Ko, Seongho Park, Jihee Kim, and Juho Kim. 2020. Solutionchat: Real-time moderator support for chat-based structured discussion. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.

[27] Ian Li, Anind K Dey, and Jodi Forlizzi. 2011. Understanding my data, myself: supporting self-reflection with ubicomp technologies. In *Proceedings of the 13th international conference on Ubiquitous computing*. 405–414.

[28] Michael A Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. 2020. Co-designing checklists to understand organizational challenges and opportunities around fairness in AI. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu HI USA). ACM, New York, NY, USA.

[29] David Madras, Toniann Pitassi, and Richard Zemel. 2017. Predict responsibly: improving fairness and accuracy by learning to defer. *arXiv preprint arXiv:1711.06664* (2017).

[30] MICHAEL MANKINS, ERIC GARTON, and DAN SCHWARTZ. 2021. Future-Proofing Your Organization. (cover story). *Harvard Business Review* 99, 5 (2021), 42 – 48. http://ezproxy.lib.utexas.edu/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=151926545&site=ehost-live

[31] Tariq Masood, Duncan McFarlane, Jennifer Schooling, and Ajith Parlikad. 2015. The role of futureproofing in the management of infrastructural assets. UCL STEaPP.

[32] Laura Matrajt, Julia Eaton, Tiffany Leung, and Elizabeth R Brown. 2021. Vaccine optimization for COVID-19: Who to vaccinate first? *Science Advances* 7, 6 (2021), eabf1374.

[33] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)* 54, 6 (2021), 1–35.

[34] Michael Quinn Patton. 1990. *Qualitative evaluation and research methods*. SAGE Publications, inc.

[35] Inioluwa Deborah Raji, Andrew Smart, Rebecca N White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 33–44.

[36] BRIAN D Rich. 2014. Principles of future proofing: a broader understanding of resiliency in the historic built environment. *Preserv Educ Res* 7 (2014), 31–49.

[37] Devansh Saxena, Karla Badillo-Urquiola, Pamela J Wisniewski, and Shion Guha. 2021. A framework of high-stakes algorithmic decision-making for the public sector developed through a case study of child-welfare. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–41.

[38] Mike Schaekermann, Graeme Beaton, Minahz Habib, Andrew Lim, Kate Larson, and Edith Law. 2019. Understanding expert disagreement in medical data analysis through structured adjudication. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–23.

[39] Mike Schaekermann, Joslin Goh, Kate Larson, and Edith Law. 2018. Resolvable vs. irresolvable disagreement: A study on worker deliberation in crowd work. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–19.

[40] Hong Shen, Haojian Jin, Ángel Alexander Cabrera, Adam Perer, Haiyi Zhu, and Jason I Hong. 2020. Designing Alternative Representations of Confusion Matrices to Support Non-Expert Public Understanding of Algorithm Performance. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–22.

[41] Rachel Emma Silverman and Nikki Waller. 2015. The algorithm that tells the boss who might quit. https://www.wsj.com/articles/the-algorithm-that-tells-the-boss-who-might-quit-1426287935

[42] Megha Srivastava, Hoda Heidari, and Andreas Krause. 2019. Mathematical notions vs. human perception of fairness: A descriptive approach to fairness for machine learning. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2459–2468.

[43] Halil Toros and Daniel Flaming. 2018. Prioritizing homeless assistance using predictive algorithms: an evidence-based approach. *Cityscape* 20, 1 (2018), 117–146.

[44] Niels Van Berkel, Jorge Goncalves, Danula Hettiachchi, Senuri Wijenayake, Ryan M Kelly, and Vassilis Kostakos. 2019. Crowdsourcing perceptions of fair predictors for machine learning: a recidivism case study. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–21.

[45] Christo Wilson, Avijit Ghosh, Shan Jiang, Alan Mislove, Lewis Baker, Janelle Szary, Kelly Trindel, and Frida Polli. 2021. Building and auditing fair algorithms: A case study in candidate screening. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 666–677.

[46] Allison Woodruff, Sarah E Fox, Steven Rousso-Schindler, and Jeffrey Warshaw. 2018. A qualitative exploration of perceptions of algorithmic fairness. In *Proceedings of the 2018 chi conference on human factors in computing systems*. 1–14.

[47] Yao Xie, Melody Chen, David Kao, Ge Gao, and Xiang'Anthony' Chen. 2020. CheXplain: Enabling Physicians to Explore and Understand Data-Driven, AI-Enabled Medical Imaging Analysis. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[48] Qian Yang, Jina Suh, Nan-Chen Chen, and Gonzalo Ramos. 2018. Grounding interactive machine learning tool design in how non-experts actually build models. In *Proceedings of the 2018 Designing Interactive Systems Conference*. 573–584.

[49] Zining Ye, Xinran Yuan, Shaurya Gaur, Aaron Halfaker, Jodi Forlizzi, and Haiyi Zhu. 2021. Wikipedia ORES Explorer: Visualizing Trade-offs For Designing Applications With Machine Learning API. In *Designing Interactive Systems Conference 2021*. 1554–1565.

[50] Bowen Yu, Ye Yuan, Loren Terveen, Zhiwei Steven Wu, Jodi Forlizzi, and Haiyi Zhu. 2020. Keeping designers in the loop: Communicating inherent algorithmic trade-offs across multiple objectives. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference*. 1245–1257.